Deep Learning to Identify Cancer Risk in Patients

Swati Sisodia
Computer and Information Sciences and Engineering Department
University of Florida

Abstract - Cancer is fundamentally a disease of the genome, caused by changes in the DNA, RNA, and proteins of a cell that push cell growth into overdrive [1]. Identifying the genomic alterations that arise in cancer can help researchers decode how cancer develops and improve upon the diagnosis and treatment of cancers based on their distinct molecular abnormalities.

The aim of the project is to integrate micro RNA sequences(mRNA), Reverse Phase Protein Array (RPPA), Copy Number Alterations (CNA) and clinical data from 10 cancer types to train and evaluate classification models using traditional methods and deep learning.

I. INTRODUCTION

With the advent of Modern genomic technologies, vast amounts of data has been produced for healthy as well as diseased cohorts. Large repositories of genomic datasets are now available for analysis which can help in diagnosis, prognosis and treatment on a personalized basis. However there are several challenging aspects in learning models of these genomic sequences, primarily high dimensionality of data as compared to small number of samples. Irrelevant and Noisy features effects complexity as well as performance of the learning algorithms.

Traditional methods try to overcome this challenge using feature selection which helps remove the redundant and noisy features producing a small subset of features upon which learning algorithms like SVM can be easily trained. A second method performs feature selection while learning using regularization to place constraints on the data forcing some features to be sparse (lasso) and naturally perform features selection.

However, these models are typically shallow models with a single layer of feature transformations and they lack multiple layers of adaptive non-linear features. A third approach to build a classifier could be deep learning. Deep learning models exploits many layers of information processing stages in hierarchical architectures for unsupervised feature learning and for pattern analysis/classification. These models have potential of learning intricate patterns from even high-dimensional raw data with little guidance. Genome data is an ideal candidate for genome data, given its high dimensionality as well as lack of knowledge about interactions between the features.

This project aims to explore the three approaches to train, evaluate and compare classification models for different typology of tumors in the dataset.

II. RELATED WORK

There are a number of studies on genomic data that explores different aspects of data. In 2014, Zhang et al. proposed an integrative framework to identify genetic and epigenetic features related to ovarian cancer and to quantify the causal relationships among these features using a probabilistic graphical model based on the Cancer Genome Atlas (TCGA) data. The constructed Bayesian network has identified some new genetic/epigenetic pathways, and help understand better the molecular mechanisms of ovarian cancer. Kamdar obtained a set of genomic features which provided evaluation metrics for each typology through Genomic Co-occurrence-based feature selection to classify Subtypes of Breast Invasive Carcinoma.

In 2015, Chen et al. proposed a computational workflow to successfully use Reverse Phase Protein Array Profiles to classify patient samples into ten main cancer types. With 10-fold cross-validation on the training set, the SMO (Sequential minimal optimization) and the IFS (Incremental Feature Selection), methods were used to choose an optimal feature set. The methods could provide clinicians with knowledge of key distinct biochemical features of cancer types and shed some new light on the discoveries of specific biomarkers of different types of cancers.

Two recent methods, DeepBind (Alipanahi et al., 2015) and DeepSEA (Zhou and Troyanskaya, 2015), successfully applied deep learning to modeling the sequence specificity of protein binding with a performance superior to the best existing conventional learning methods.

An interesting recent research by Tan 2015 et al. introduced an unsupervised feature construction approach based on DAs that summarizes available genomic data and extracts useful features on large compendium of breast cancer gene expression data. The constructed features that distinguish tumor from normal samples, classify patients' estrogen receptor (ER) status, summarize intrinsic subtypes, and identify the activity of key transcription factors (TFs). Results suggest that constructed features from denoising autoencoders more predictive of survival than commonly used markers such as tumor grade or ER status and provide a fruitful ground for approaches that aim to reason from such data.

With success in using each of these high-throughput technologies to build patient classifiers for various phenotypes and outcomes, the next frontier explores "multi-modal" data for the same set of subjects and conduct integrative analyses using multi-level views on the same phenomena. Ray et al., 2014 applied integrative multi-modal classification technique called "Multi-modal ensemble" (MME) approaches apply methods to "ensemble" multiple classification models derived from individual data modalities. The classification of subjects is then performed by an ensemble classification model, which is defined as a function of models from individual data modalities.

III. METHODS

1 *Data Collection*

The dataset for the project is a set of 4442 patients for 10 different types of cancer published by The Cancer Genome Associations (TCGA). The raw TCGA data is available as zipped text files from their data portal. The obtained data include clinical information, mRNA gene expression, CNAs, and mRNA. The sample size of the dataset is summarized in Table 1. Data is labeled as cancer subtype in form of ICD10 codes in clinical summary of the data.

Table 1: Data Summary

Cancer Type	Sample
Brain Lower Grade Glioma	426
Breast invasive carcinoma	868
Colon Adenocarcinoma	310
Glioblastoma multiforme	66
Head and Neck squamous cell carcinoma	344
Kidney renal clear cell carcinoma	469
Lung Adenocarcinoma	359
Lung squamous cell carcinoma	323
Ovarian serous cystadenocarcinoma	219
Prostate adenocarcinoma	346
Total	3730

2 *Data Preparation and Preprocessing*

The Level 3 TCGA data provides normalized CNA , mRNA and RPPA for aggregated /segmented regions, per sample, normalized protein expression for each gene, per sample and the calculated expression signal of a gene, per sample. Only patients with complete features for all three CNA , mRNA and RPPA are retained into final dataset resulting in an intersection of 3800 patients from a total of 5944 patients. As the missing rate is relatively low, simple imputation was used using median values across samples for fields.

3. *Model Development*

3.1 Feature Selection: There are > 20000 protein-coding genes in the human genome along with copy number alterations and, other auxiliary genomic segments, which increases our feature space drastically (~ 45000 features). To reduce the dimensionality and filter out unrelated feature, traditional L1-based feature selection was used, which examines each feature individually to determine the strength of the relationship of the feature with the response variable. Selected features were to train the classification models. Top 20 features are summarized in table 1 in supplements.

3.2 Classification Methods:

Four different models have been evaluated for the classification task.

SVM:

This includes the commonly used soft-margin SVM with L1-loss (for one or two classes, with regularized or no offset, with or without using a kernel). Here the matrix A ∈ R d×n contains all n data points as its columns, and Δ is the unit simplex in R n, being the set of non-negative vectors summing up to one (i.e. probability vectors).

$$\min_{x \in \triangle} \|Ax\|_2^2$$

The data was split into train and test using 90:10 split. Margin for the training the model was determined using 5 fold cross validation on train data using grid search. The optimum Margin value came out to be 100 which was used to train the final model and tested on the test set.

Random Forests:

Random Forest is a generic principle of classifier combination that uses L tree-structured base classifiers {h(X,Θn), N=1,2,3,…L}, where X denotes the input data and {Θn} is a family of identical and dependent distributed random vectors. Every Decision Tree is made by randomly selecting the data from the available data and the features are randomly selected in each decision split.

The data was split into train and test using 90:10 split. Optimum number of tress as well the size of subset of features used at each split for the training the model was decided using 5 fold cross validation on train data using grid search. Test set was then evaluate on best performing model and report the final accuracy.

Lasso:

Lasso is given by constrained variant of L1-regularized least squares regression. Here the right-hand side b is a fixed vector b ∈ Rd, and ◊ is the L1-unit-ball in Rn. If the desired L1-regularization constraint is not ||x||1 ≤ 1, but ||x||1 ≤ r for some r > 0 instead, then it is enough to simply re-scale the input matrix A by a factor of (1/r).

$$\min_{x \in \blacklozenge} \|Ax - b\|_2^2$$

The train and test split was 90:10. The optimal alpha was determined from the train set by plotting coefficient error as a function of regularization and select most

Stacked Denoising Autoencoders (SDAE):

Denoising auto-encoder is a stochastic version of the auto-encoder which tries to derive a more robust network by trying to undo effect of noise. When stacked to form a deep network by feeding the latent representation of the denoising autoencoder found on the layer below as input to the current layer. Once all layers are pre-trained, the network goes through a second stage of training is fine-tuned to minimize prediction error on a supervised task.

The data was split into train and test using 90:10 split. A grid search was performed across parameters with cross entropy as loss function. Cross validation was not performed due to large computational time but each model was tested on the test set and best model is reported in the results section.

The process flow highlighting the details of the implementation is shown in the fig. 1 in supplements. The parameters across which grid search was performed for different models are summarized in table 2 in supplements.

IV. RESULTS

The accuracy, precision, and recall were calculated for each of the learning models. Table 2 shows all the parameters of each model. Table 3 summarizes the results.

Table 2: Model Parameters

Model	Parameters
SVM	C = 100
RF	Number of trees = 40, Number of features = 10
Lasso	Alpha = 0.0001
SDAE	Epochs= 100, Learning rate = 0.01, batch size =20, Hidden Layer Size = [1000,1000,1000], noise =0.2

Table 3: Evaluation Metrics for Classifiers

Model	Accuracy	Precision	Recall
SVM	0.94638	0.95	0.95
Lasso	0.6	0.75	0.60
Random Forests	0.9544	0.95	0.95
SDAE	0.4666	0.47	0.30

It can be seen that the SVM and the RF classifiers have minimal Test errors (< 5.0%), and a very high precision and recall (> 97.0%) values across all typologies. The performance of Lasso Classifier was considerably poorer than SVM and RF with average precision and recall. Stacked Denoising Autoencoders gave worst performance with accuracy less than 20% and 0.03 precision.

V. DISCUSSIONS AND CONCLUSIONS

The goal of the project was to analyze multimodal data using deep learning. Simple models with Feature selection combined with SVM and RF give quite good performance.

Lasso which usually works well with high dimensional data gives a lower performance in comparison. However, accuracy improved from 47 percent to 60 percent when Lasso was performed on a trimmed dataset with 200 instances from each class suggesting sensitivity of lasso to balanced data.

SDAE does not perform across entire grid search of parameters summarized in table 2 in supplements. For every possible combination in the grid, most samples are classified to same two or three classes suggesting that model is too complex for the classification task for the given dataset or number of samples is too less for SDAE to learn meaningful features. Like Lasso , there was a considerable improvement in performance a more balanced dataset was used.

Future work can involve more exhaustive grid search to test if SDAE can perform better possibly with more number of samples and a more balanced dataset.

Secondly, different TCGA molecular datasets (SNP, DNA methylation) can be evaluated individually and then in combination to construct derive more useful abstractions.

VI. REFERENCES

[1] gdc-portal.nci.nih.gov. N.p., 2016. Web. 19 Sept.

2016.

[2] Shingo Tsuji, Hiroyuki Aburatani. Deep learning for the large-scale cancer data analysis. [abstract]. In:Proceedings of the AACR Special Conference on Computational and Systems Biology of Cancer; Feb 8-11 2015; San Francisco, CA. Philadelphia (PA):AACR; Cancer Res 2015;75(22 Suppl 2):Abstract nr B1-08.

[3] Gene expression inference with deep learning.Yifei Chen, Yi Li, Rajiv Narayan, Aravind Subramanian, Xiaohui Xie bioRxiv 034421; doi:http://dx.doi.org/10.1101/034421

[4] Tan J, Ung M, Cheng C, Greene CS.Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. Pac Symp Biocomput.2015;:132-43.

[5] Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. Molecular Systems Biology. 2016;12(7):878. doi:10.15252/msb.20156651.

[6] Zhang Q, Burdette JE, Wang JP. Integrative network analysis of TCGA data for ovarian cancer.BMC Syst Biol. 2014;8:1338.

[7] Kamdar, Maulik R. Visualizing Personalized Cancer Risk Prediction. 1st ed. Web. 22 Sept. 2016.

[8] Deng, Li. "Three classes of deep learningarchitectures and their applications: a tutorial survey."APSIPA transactions on signal and information processing (2012).

[9] Zhao Q, Shi X, Xie Y, Huang J, Shia B, Ma S. Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. *Briefings in Bioinformatics*. 2015;16(2):291-303. doi:10.1093/bib/bbu003.

[10] Classifying Ten Types of Major Cancers Based on Reverse Phase Protein Array Profiles Pei-Wei Zhang, Lei Chen, Tao Huang , Ning Zhang , Xiang-Yin Kong , Yu-Dong Cai ; March 30, 2015;http://dx.doi.org/10.1371/journal.pone.0123147

[11] Practical recommendations for gradient-based training of deep architectures, Yoshua Bengio, U. Montreal, arXiv report:1206.5533, Lecture Notes in Computer Science Volume 7700, Neural Networks: Tricks of the Trade Second Edition, Editors: Grégoire Montavon, Geneviève B. Orr, Klaus-Robert Müller, 2012.

[12] ADAGE analysis of publicly available gene expression data collections illuminates Pseudomonas aeruginosa-host interactions Jie Tan, John H Hammond, Deborah A Hogan, Casey S Greene bioRxiv 030650; doi: http://dx.doi.org/10.1101/030650

[13] Applications of Deep Learning in Biomedicine Polina Mamoshina, Armando Vieira, Evgeny Putin, and Alex Zhavoronkov Molecular Pharmaceutics 2016 13 (5), 1445-1454 DOI: 10.1021/acs.molpharmaceut.5b00982

[14] Haoyang Zeng, Matthew D. Edwards, Ge Liu, David K. Gifford; Convolutional neural network architectures for predicting DNA–protein binding. Bioinformatics 2016; 32 (12): i121-i127. doi: 10.1093/bioinformatics/btw255

[15] Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. In: Advances in Neural Information Processing Systems, vol. 19, p 153 (2007)

[16] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research, 11:3371–3408, 2010.

[17] R. Sager, “Expression genetics in cancer: shifting the focus from DNA to RNA,” Proceedings of the National Academy of Sciences of the United States of America, vol. 94, no. 3, pp. 952–955, 1997.

[18] Ray, B., Henaff, M., Ma, S., Efstathiadis, E., Peskin, E.R., Picone, M., Poli, T., Aliferis, C.F., Statnikov, A., 2014. Information content and analysis methods for multi-modal high-throughput biomedical data. Sci. Rep. 4, 44